2010-01-01

Inside the Selection Box: Visualising active learning selection strategies

Brian Mac Namee
Technological University Dublin, brian.macnamee@tudublin.ie

Rong Hu
Technological University Dublin

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Recommended Citation

MacNamee,B., Hu,R.,& Delany, S. (2010) Inside the Selection Box: Visualising active learning selection strategies. *The Challenges of Data Visualization Neural Information Processing Systems*, (NIPS) 2010 Workshop, Vancouver, Canada, 6 December. doi:10.21427/D79P5S

Inside the Selection Box: Visualising active learning selection strategies

Brian Mac Namee, Rong Hu & Sarah Jane Delany
Applied Intelligence Research Centre
Dublin Institute of Technology
Dublin 8, Ireland
Brian.MacNamee@dit.ie

Abstract

Visualisations can be used to provide developers with insights into the inner workings of interactive machine learning techniques. In active learning, an inherently interactive machine learning technique, the design of selection strategies is the key research question and this paper demonstrates how spring model based visualisations can be used to provide insight into the precise operation of various selection strategies. Using sample datasets, this paper provides detailed examples of the differences between a range of selection strategies.

1 Introduction

In order to reach better solutions, some machine learning algorithms can benefit from the guidance of a human analyst. Commonly referred to as *interactive* (or *human-in-the-loop*) machine learning [1], such approaches mix the ability of automated algorithms to deal with massive amounts of multivariate data, with our own ability to identify complex patterns. One of the best ways in which to enable the interaction between algorithms and analysts is to use visualisations of the underlying datasets.

Active learning (AL) [2] is a semi-supervised machine learning approach that is inherently interactive. The goal of AL is to overcome the problem in supervised learning that labelled datasets can be difficult or expensive to obtain by training a classifier from an unlabelled dataset by asking an analyst to label a small number of examples chosen by an automated *selection strategy* to be most informative. While there is potential to use visualisations to help analysts guide the AL process itself, there is also potential to use visualisations to help design AL algorithms. Probably the most important research challenge in developing AL systems is selecting the most appropriate selection strategy. Typically, selection strategies are compared by examining learning curves that indicate the accuracy of the classifier created after various numbers of labels have been solicited from the analyst and used to build a classifier [3]. This, however, does not provide any explanation as to why certain selection strategies work and others do not. There is potential to use visualisations to provide this explanation.

In this paper we present the *Case Base Topology Viewer for Active Learning* (CBTV-AL) a system designed to visualise the AL process so that the operation of selection strategies can be better understood. In Section 2 we provide a brief overview of related work in interactive machine learning and visualisation, before explaining the CBTV-AL system in Section 3. In Section 4 we demonstrate how CBTV-AL can be used to visualise the AL process using two classification datasets, and describe how this offers insight into the operation of different selection strategies. Finally, in Section 5 we conclude and outline the directions in which we intend to take this work in the future.

2 Related work

Interactive machine learning [1] requires human analysts to contribute to semi-automated machine learning processes in order to reach better solutions. When applied to the correct task, interactive machine learning can achieve results that fully automated solutions might never reach. For example, the *iVibrate* [4] system presents users with intermediate clustering results and asks them to refine the cluster boundaries found by the algorithm before the clustering process continues. In image retrieval (e.g. [5]) an initial set of images returned from a textual search query can be presented to a user and this set can be revised and extended based on images, or subsets of images, selected by the user as most relevant. Even the process of building classifiers can benefit from interactive user input on which features to use, the relative importance of training examples or values for model parameters [6, 1].

One of the most effective ways that interactions between machine learning techniques and users can be facilitated is to use visualisations of the dataset being used [7, 6, 8, 1]. For example, in the iVibrate system [4] a two dimensional scatter plot of the data being clustered is presented to users to allow them refine the cluster boundaries that have been found by the algorithm so far. There are a wide range of approaches to visualising large, high-dimensional datasets [9] but it is essentially a dimensionality reduction exercise. Examples include the Grand Tour [10], principal components analysis [11], force-directed graph drawing algorithms [12], Sammon mapping [13], and techniques that make use of output of the machine learning process itself [14].

There are few examples of visualisation applied to active learning. Visalix [15] is one, the core of which is a 3D interactive visual clustering component. It also provides a visual active learning component which visualises the data in an uncertainty space and allows the user to select the next example to be labelled. It is limited in its applicability as the visualisation is built on representations of the dataset examples based on the certainty with which the examples can be classified. This does not allow for the visualisation of the existing automated selection strategies that have proven successful in active learning, or allow users to guide these selection strategies by informed manual selection. To the best of our knowledge there are no previous attempts to use visualisations to better understand the workings of active learning selection strategies. The next section will describe our approach to this.

3 Generating Visualisations for Active Learning

The active learning (AL) process begins with an unlabelled dataset and so visualisation techniques that do not rely on classifier output are most suitable. In CBTV-AL we use the force-directed graph drawing algorithm known as the *spring model* [12] which allows the display of n-dimensional data on a two dimensional plane by using the *similarity* between examples to dictate their relative positions on the graph. Amongst other things this approach has been used to show the impact of adding or removing examples from a dataset [16] and the effect of using different measures of the similarity between examples in a dataset [17]. There are many ways to measure the similarity between examples [18]. In our system we use two of the most common – normalised Euclidean distance for datasets where all features are numeric and cosine similarity for textual datasets for which we use a *bag-of-words* representation [19]. More details of our approach to dataset visualisation can be found in [17].

Before the AL process begins the spring model graph of a dataset is allowed to reach equilibrium. The first step in the AL process is the selection of a small (typically 5 – 10) set of examples to form an *initial training set* which is used to seed the AL process. In our system the initial training set can be selected either at random or using agglomerative hierarchical clustering (AHC) clustering [20]. Using this initial labelled set all examples are ranked in order of their useful to the AL process. The method for calculating this ranking depends on the selection strategy being used. The selection strategy is used to select a number of examples from the dataset (usually one) for labelling by a human oracle. We consider four selection strategies:

Density sampling in which the density [21] of each example in a dataset is calculated as the sum of the similarities of examples within a pre-defined region around the target example, and examples in the most dense regions of the dataset are selected first for labelling.

Density & diversity sampling which selects examples for labelling based a weighted combination of density and diversity to create a more balanced selection strategy [21].

Uncertainty sampling in which the examples that a classifier, trained on the examples labelled by the oracle so far, has most difficulty classifying are considered most uncertain and selected for labelling first [22, 23].

Once the oracle has labelled new examples the remaining examples in the dataset is are re-ranked according to their usefulness to the AL process (in our system this involves recalculating diversity and uncertainty as density remains constant throughout the process). The process repeats until some stopping criteria is reached (typically a labelling budget expires). The purpose of CBTV-AL is to visualise this process. This is done by showing a graph of the dataset, arranged using the spring model, and annotating this graph to display labels given by the analyst, predictions made using a classifier built from the current dataset and measures of density, diversity and uncertainty. Section 4 will present examples of this using two classification datasets.

4 Examples of Visualised AL

The first dataset we use to demonstrate the CBTV-AL system is the popular Iris dataset (available from the UCI Machine Learning repository [24]) which has 3 classes, 4 numeric features and 150 instances (50 of each class). While the Iris dataset represents a very easy classification problem it is useful as it facilitates explanation of the AL processes and the impact of different selection strategies. The second dataset used is a binary classification dataset generated from the Reuters collection[1] and including 250 randomly sampled documents from each of the *acq* and *earn* categories. Texts are tokenised at the word level and feature values are recorded as unit length normalised word frequencies. Stop-word removal and document frequency reduction (removing all words that occur in less than 3 documents in the dataset) was also performed resulting in 1,000 features. The Reuters dataset represents a more difficult classification problem and is of much higher dimension than the Iris dataset making it a more interesting visualisation proposition. Both of these datasets are fully labelled and so the AL process is simulated removing the need for a human oracle. The purpose of these demonstrations is not to evaluate the active learning process itself, but rather to demonstrate how visualisations can be effective in understanding the inner workings of various selection strategies, justifying our use of this simulation strategy.

Figure 1 shows snapshots of the formation of the spring model graph for the Iris dataset in which examples form into three noticeable clusters. Once the graph has reached equilibrium, an initial training set of six examples are randomly selected from the full dataset. Figure 2(a) shows the initial labelled examples enlarged, where both the colours and shapes of these labelled examples represent their *true class*. A *5-nearest neighbour* classifier using distance-weighted majority voting and built from this initial training set is used to generate a *predicted class* for each example remaining in the unlabelled set. The predictions made by this classifier are indicated by colour in Figure 2(a), where the shapes of each example represent true class. It is clear that the classifier is doing a particularly bad job in this case, particularly for the class represented by circles (this will be explained below). Figure 2(b) shows the same dataset where the darkness of each unlabelled example indicates its density (darkness increases with density). It is evident from the graph that examples are most dense in the centre region. Figure 2(c) shows the initial diversity scores for the examples in the dataset (example darkness increases with diversity) and the relationship between diversity and distance to a labelled example is evident. Figure 2(d) shows the initial classification uncertainty ratings for each unlabelled example (darkness increases with classification uncertainty). It is interesting to compare Figures 2(d) and 2(a) to note that at this stage classifications that are quite certain are not always accurate.

Figure 3 shows a series of snapshots of the results of the AL process running on the Iris dataset using a density-only selection strategy (it is worth noting again that density ratings remain constant throughout the AL process). While density-based selection strategies are intuitively sensible, the ineffectiveness of density-only strategies is clearly evident as one class (illustrated using red stars) is more densely packed than the other two so all labelling is initially done for examples of that class.

[1] Available at: http://www.daviddlewis.com/resources/testcollections/reuters21578/

It is only after this class is almost completely labelled that examples from either of the other classes are presented for labelling, and the performance of the resulting classifier built improves.

Figure 4 shows snapshots of the same process using a selection strategy based on a combination of density and diversity. This approach achieves a much more successful balance between exploration and exploitation and examples from all over the graph are chosen for labelling by the oracle. This leads to very accurate classifiers very early on in the process – indicated by the agreement between colour and shape of unlabelled examples.

Figure 5 shows snapshots of the same process using classifier uncertainty sampling selection strategy. Uncertainty sampling concentrates on the class boundaries and it is interesting to compare Figures 5(d) and 4(d) to note how the density & diversity selection moves around the full example space while uncertainty sampling concentrates on the class boundaries - in particular the boundary between the star and triangular classes. The initial training set used in this process was selected randomly and the impact of this can be seen in this example. Figure 5(b) shows that a number of steps into the AL process the predicted classes, particularly for the class represented by circles, are not accurate, especially when compared to Figure 4(b). Because the classifier built from this set is not very successful the uncertainty scores do not serve as a good guide to selection. This clearly illustrates the need in active learning to make an informed selection of the initial training set (using techniques such as clustering [25]) when uncertainty sampling selection strategies are being used.

Figures 2(e) and 2(f) show the final class labellings and final classification uncertainties (calculated by performing a *leave-one-out* cross validation on the fully labelled dataset) for the Iris dataset. It is interesting to note how clearly the visualisation indicates that classification uncertainty is centred on the border between the two non-linearly separable classes, while examples from the third linearly separable class have high classification certainties associated with them.

Figure 6 shows the formation of the visualisation graph for the Reuters dataset and Figure 7 shows the initial density, diversity, and uncertainty ratings for examples in the dataset after the initial training set has been selected. This time instead of using a random selection of the initial training set, agglomerative hierarchical clustering is used which leads to a better initial training set. This time the initial classifications made by the classifier built using the initial training set are reasonably accurate which has a significant impact on the uncertainty sampling strategy. Figure 8 further emphasises the futility of density sampling selection strategies as the entirety of one dense class is labelled before considering any examples from the second less dense class. The density and diversity sampling approach shown in Figure 9 is a major improvement showing a much more balanced exploration of the example space. Finally, Figure 10 shows the uncertainty sampling selection strategy. This time, because of the more balanced initial training set performance is considerably better than was the case for the Iris dataset in the early stages. It is also particularly interesting to compare the selection path taken by the uncertainty sampling selection strategy and the density & diversity sampling selection strategy (e.g. Figures 9(c) and 10(c)). This time selection is concentrated around the class boundaries with very few examples selected from the relatively certain region in which the members of the red class are clustered.

5 Conclusion & Future Work

The contribution of this paper is a demonstration of how visualisation techniques can be used to offer insight into the inner workings of interactive machine learning algorithms. By creating a spring model graph of a dataset and annotating it correctly, the subtleties of various selection strategies can be illustrated in order to help developers create more effective AL systems. We have used two datasets to demonstrate the differences between three common selection strategies. We intend to continue this work in two directions. The first is to use these insights to develop novel selection strategies using extra information such as data profiling information [26]. The second is to continue investigating how visualisations can be used to facilitate interactions between analysts and the active learning process – for example to create analyst-driven selection strategies.

Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/RFP/CMSF718.

References

[1] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H. Witten. Interactive machine learning: letting users build classifiers. *Int. J. Hum.-Comput. Stud.*, 55(3):281–292, 2001.

[2] Simon Tong. *Active Learning:Theory and applications*. PhD thesis, Computer science department, Stanford University, August 2001.

[3] G. Dror I. Guyon, G. Cawley and V. Lemaire. Design and analysis of the WCCI 2010 active learning challenge. In *Proc. of IJCNN-2010*, 2010.

[4] Keke Chen and Ling Liu. ivibrate: Interactive visualization-based framework for clustering large datasets. *ACM Trans. Inf. Syst.*, 24(2):245–294, 2006.

[5] James Fogarty, Desney S. Tan, Ashish Kapoor, and Simon A. J. Winder. Cueflik: interactive concept learning in image search. In *Conference on Human Factors in Computing Systems (CHI 2008)*, pages 29–38, 2008.

[6] François Poulet. Towards effective visual data mining with cooperative approaches. In *Visual Data Mining*, pages 389–406. 2008.

[7] Enrico Bertini and Denis Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *SIGKDD Explor. Newsl.*, 11(2):9–18, 2009.

[8] B. Schneiderman. Inventing discovery tools: Combining information visualization with data mining. *Information Visualization*, 1(1):5–12, 2002.

[9] Maria Cristina Ferreira de Oliveira and Haim Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE Trans. Vis. Comput. Graph.*, 9(3):378–394, 2003.

[10] D Asimov. The grand tour: A tool for viewing multidimensional data. *SIAM Journal of Scientific and Statistical Computing*, 6(1):128–143, 1985.

[11] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

[12] P. Eades. A heuristic for graph drawing. *Congressus Nutnerantiunt*, 42:149–160, 1984.

[13] J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, 1969.

[14] Petri Kontkanen, Jussi Lahtinen, Petri Myllymäki, Tomi Silander, and Henry Tirri. Supervised model-based visualization of high-dimensional data. *Intelligent Data Analysis*, 4(3-4):213–227, 2000.

[15] Loïc Lecerf and Boris Chidlovskii. Visalix: A web application for visual data analysis and clustering. In *Procs of Demonstrations Track of the 15th ACM SIGKDD*, 2009.

[16] Elizabeth McKenna and Barry Smyth. An interactive visualisation tool for case-based reasoners. *Applied Intelligence*, 14(1):95–114, 2001.

[17] Brian Mac Namee and Sarah Jane Delany. CBTV: Visualising case bases for similarity measure design and selection. In *ICCBR-10*, 2010.

[18] Padraig Cunningham. A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering*, 21:1532–1543, 2008.

[19] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[20] Ellen M. Voorhees. *The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval*. PhD thesis, Cornell University, 1986.

[21] Rong Hu, Sarah Jane Delany, and Brian Mac Namee. EGAL: Exploration guided active learning for tcbr. In *ICCBR-10*, 2010.

[22] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Procs of SIGIR-94*, pages 3–12. Springer Verlag, Heidelberg, DE, 1994.

[23] Rong Hu, Brian Mac Namee, and Sarah Jane Delany. Sweetening the dataset: Using active learning to label unlabelled datasets. In *Proceedings of the the 19TH Irish Conference on Artificial Intelligence and Cognitive Science (AICS '08)*, 2008.

[24] A. Frank and A. Asuncion. UCI machine learning repository.

[25] Rong Hu, Brian Mac Namee, and Sarah Jane Delany. Off to a good start: Using clustering to select the initial training set in active learning. In *FLAIRS 2010*, pages 26–31, 2010.

[26] Sarah Jane Delany. The good, the bad and the incorrectly classified: Profiling cases for case-base editing. In Lorraine McGinty and David C. Wilson, editors, *ICCBR-09*, volume 5650 of *LNCS*, pages 135–149. Springer, 2009.

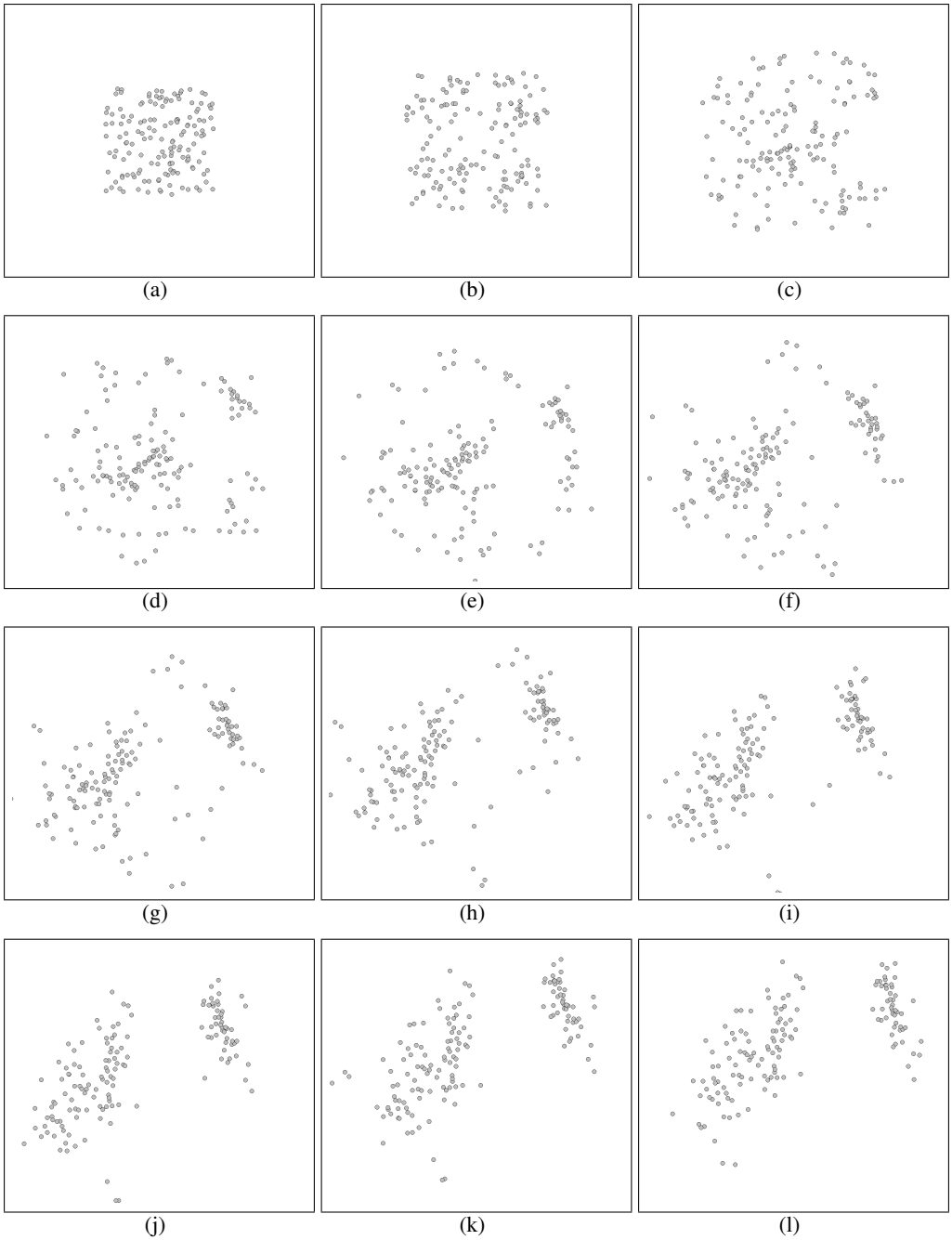
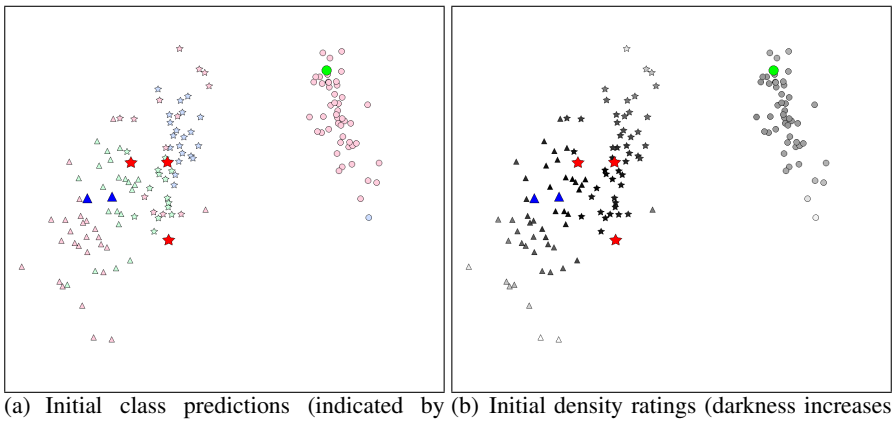
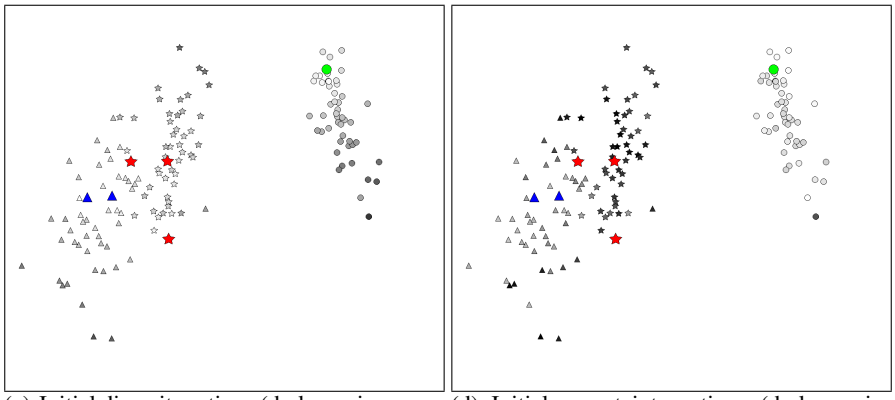


Figure 1: The spring model graph drawing process for the Iris dataset



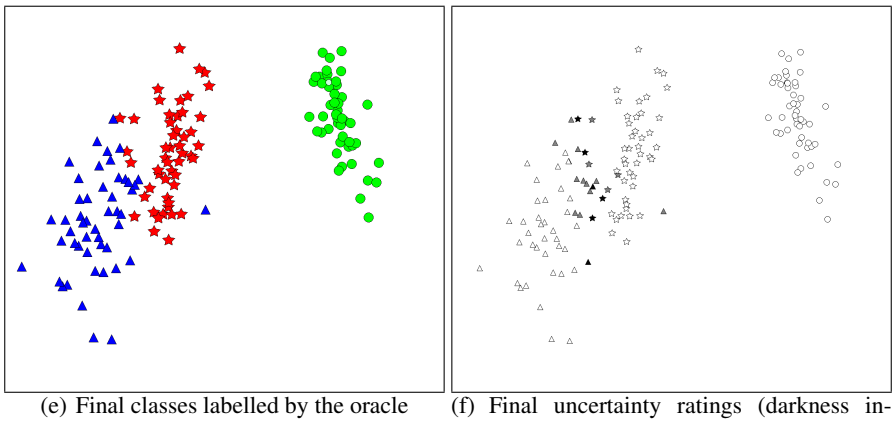
(a) Initial class predictions (indicated by colour)

(b) Initial density ratings (darkness increases with density)



(c) Initial diversity ratings (darkness increases with diversity)

(d) Initial uncertainty ratings (darkness increases with uncertainty)



(e) Final classes labelled by the oracle

(f) Final uncertainty ratings (darkness increases with uncertainty)

Figure 2: Initial class predictions, densities, diversities and uncertainties and final labelled classes and leave-one-out uncertainties for the Iris dataset. Examples labelled by the oracle are shown enlarged and their colours indicate their labelled class.

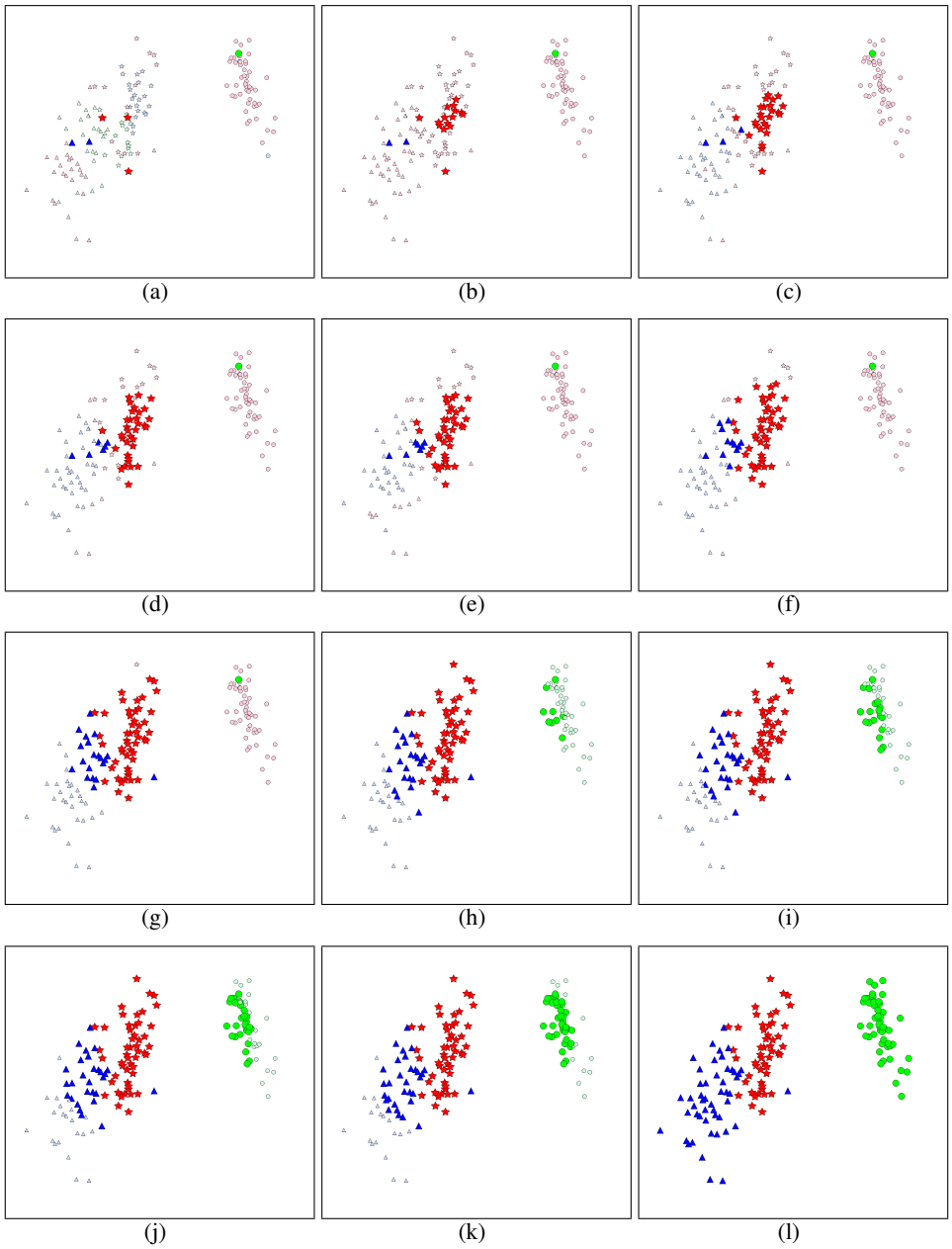


Figure 3: A visualisation of the active learning process running on the Iris dataset using a density-only selection strategy. Examples labelled by the oracle are shown enlarged. The shape of each point represents its true class. Colour indicates the true class for examples labelled by the oracle and current predicted class for those examples not yet labelled.

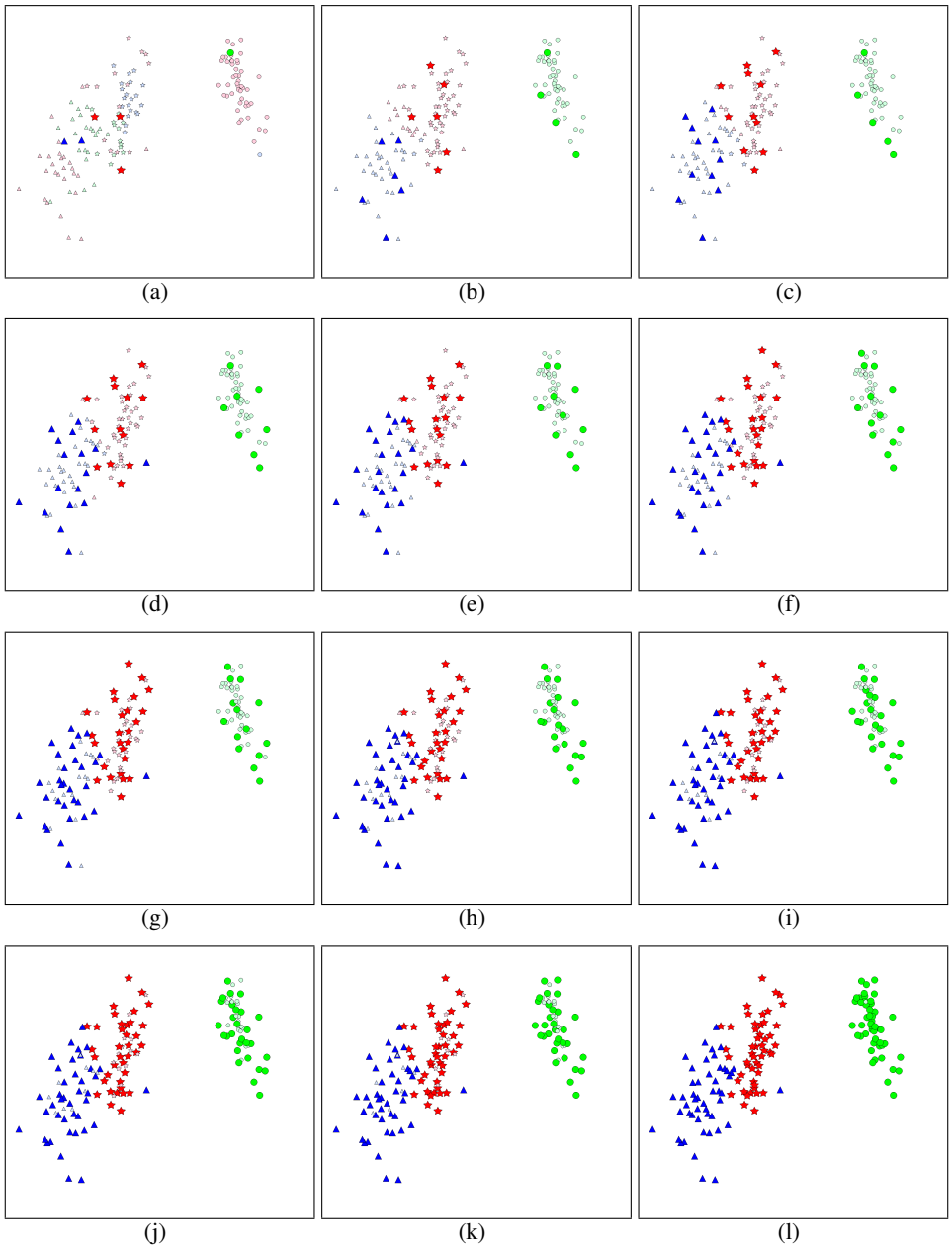


Figure 4: A visualisation of the active learning process running on the Iris dataset using a density & diversity selection strategy. Examples labelled by the oracle are shown enlarged. The shape of each point represents its true class. Colour indicates the true class for examples labelled by the oracle and current predicted class for those examples not yet labelled.

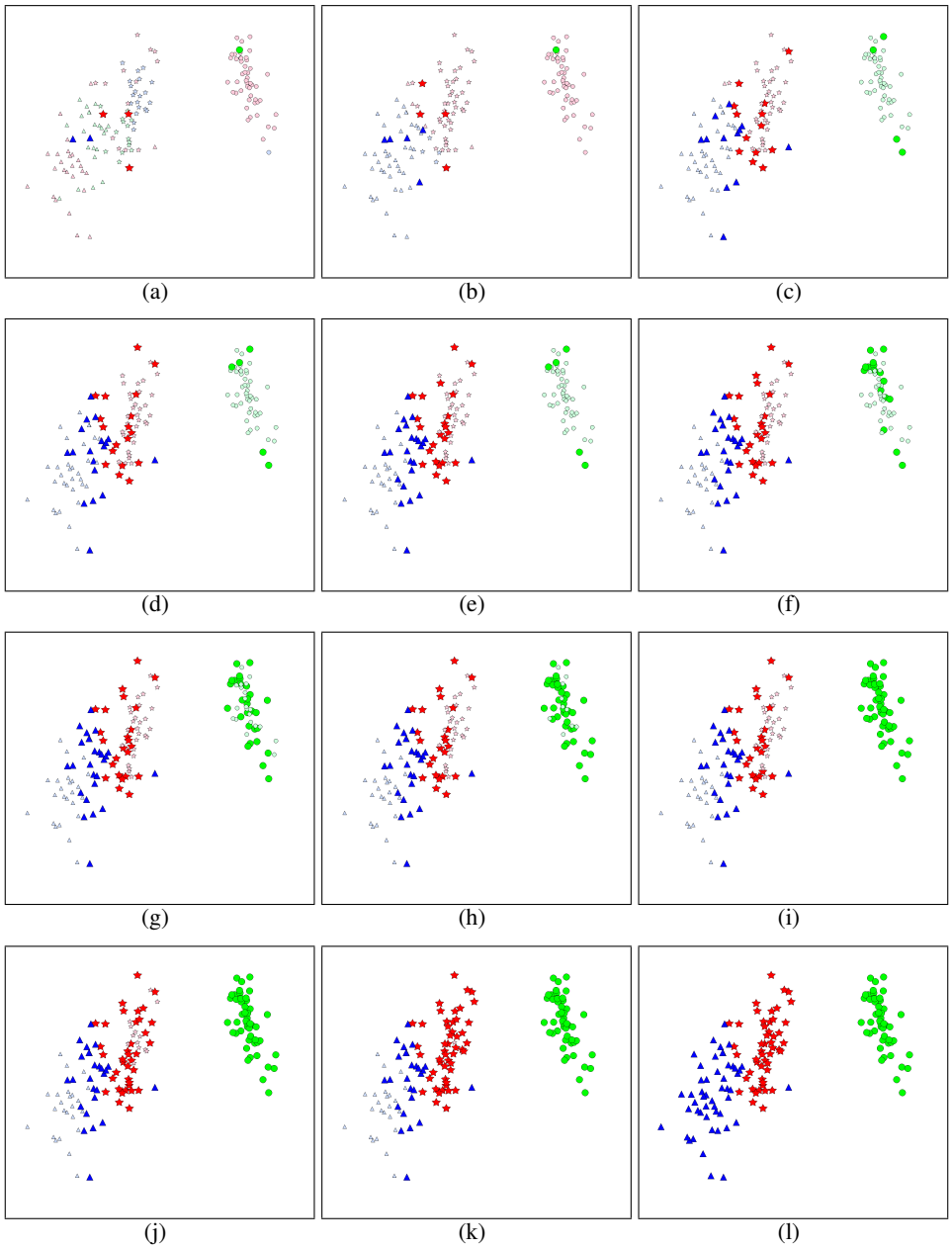


Figure 5: A visualisation of the active learning process running on the Iris dataset using an uncertainty sampling selection strategy. Examples labelled by the oracle are shown enlarged. The shape of each point represents its true class. Colour indicates the true class for examples labelled by the oracle and current predicted class for those examples not yet labelled.

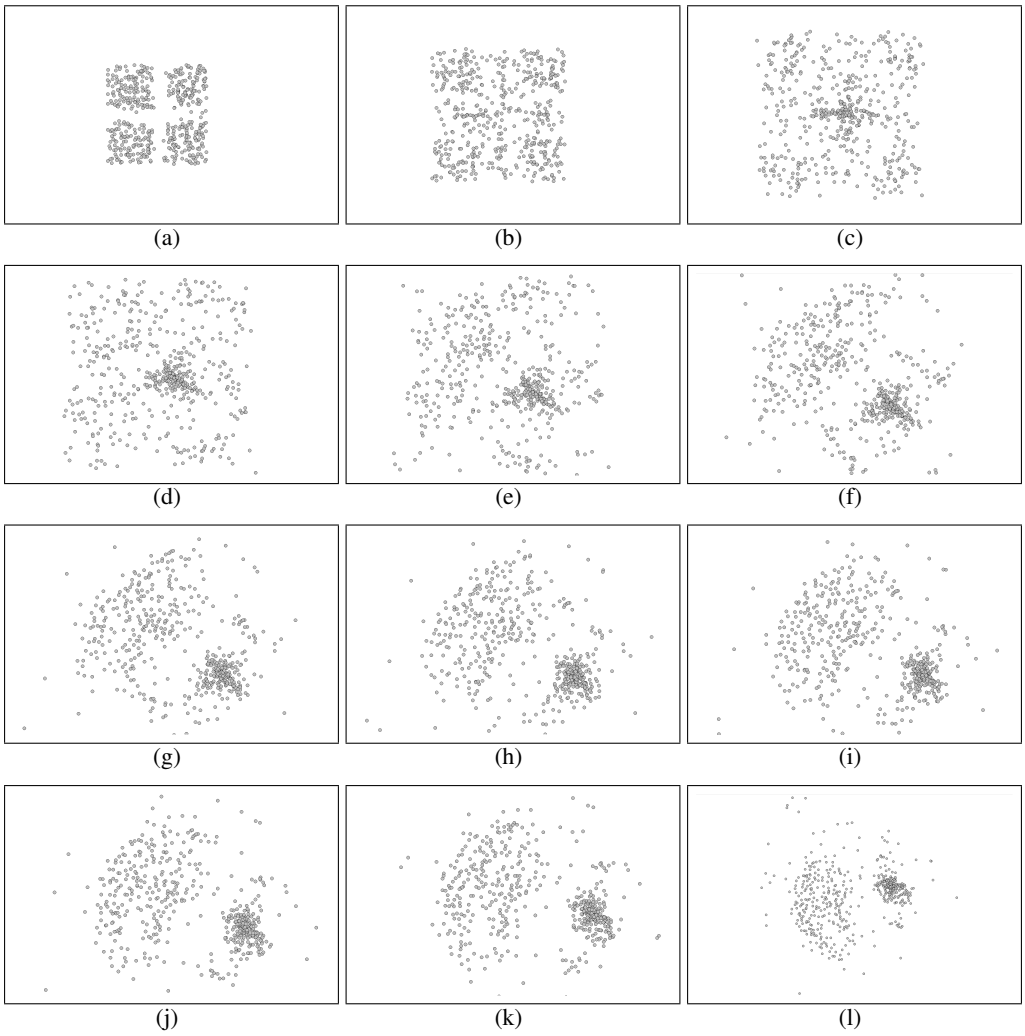
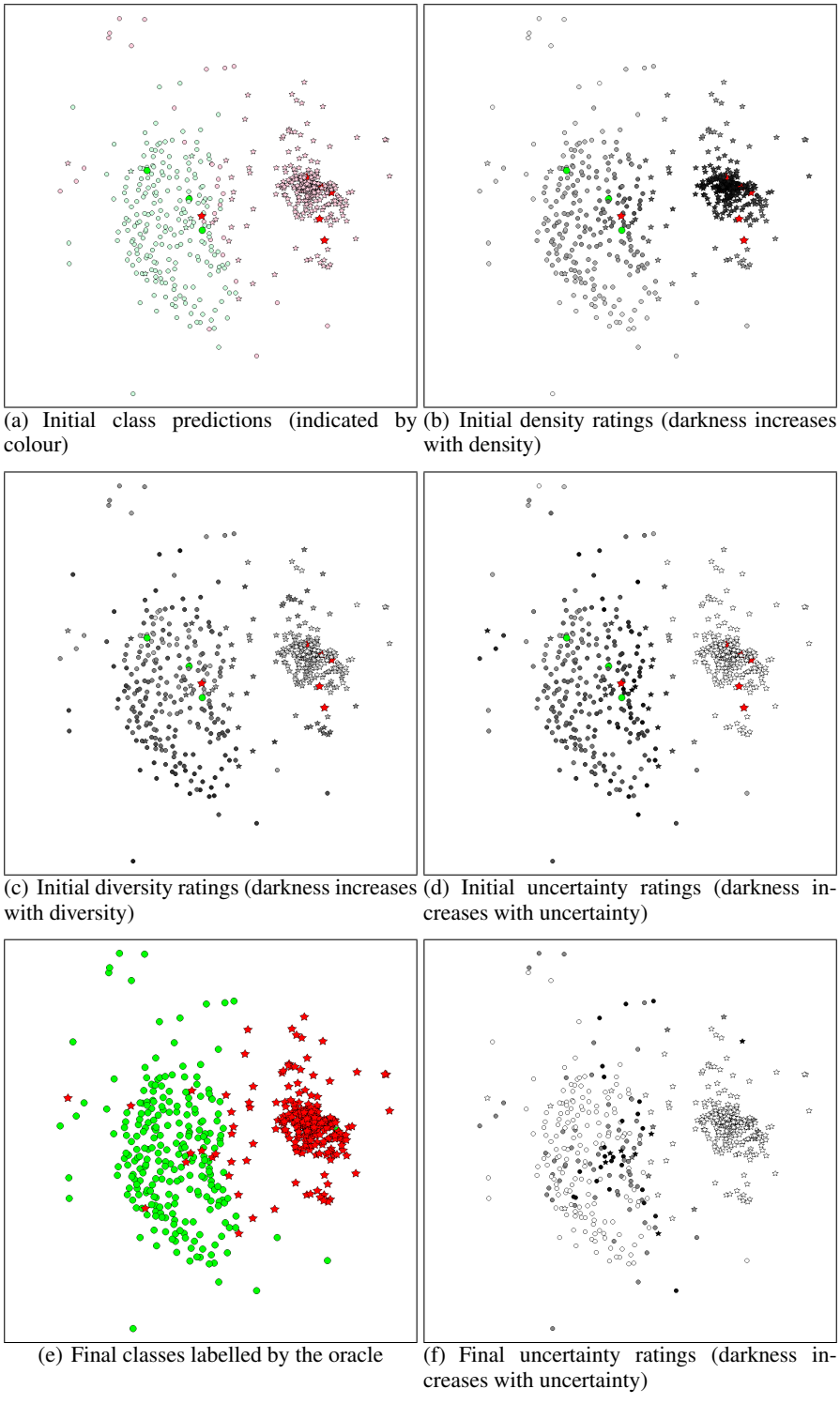


Figure 6: The spring model graph drawing process for the Reuters dataset



(a) Initial class predictions (indicated by colour)

(b) Initial density ratings (darkness increases with density)

(c) Initial diversity ratings (darkness increases with diversity)

(d) Initial uncertainty ratings (darkness increases with uncertainty)

(e) Final classes labelled by the oracle

(f) Final uncertainty ratings (darkness increases with uncertainty)

Figure 7: Initial class predictions, densities, diversities and uncertainties and final classes and leave-one-out classification certainties for the Reuters dataset.

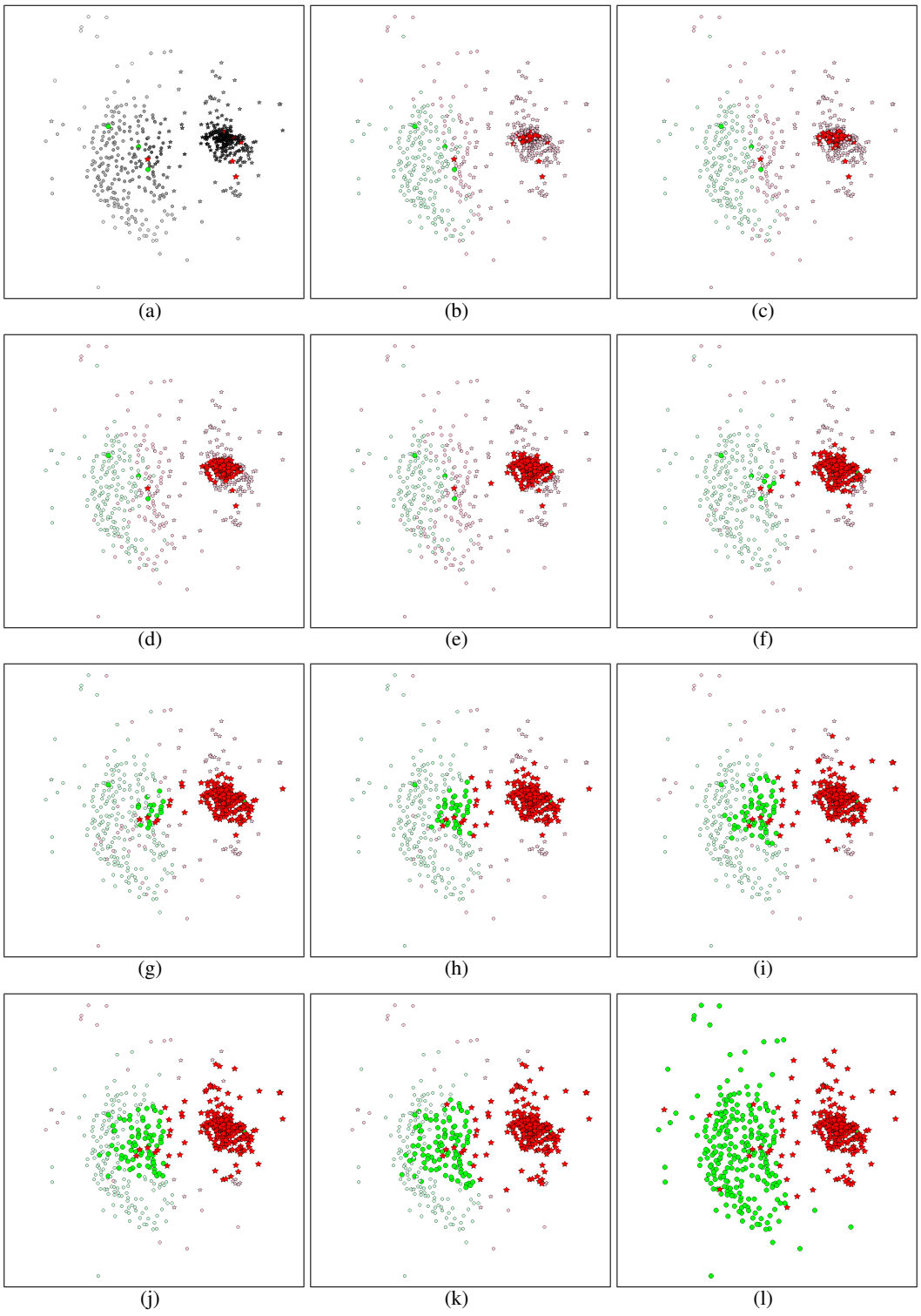


Figure 8: A visualisation of the active learning process running on the Reuters dataset using a density only selection strategy. Examples labelled by the oracle are shown enlarged. The shape of each point represents its true class. Colour indicates the true class for examples labelled by the oracle and current predicted class for those examples not yet labelled.

Figure 9: A visualisation of the active learning process running on the Reuters dataset using a density & diversity selection strategy. Examples labelled by the oracle are shown enlarged. The shape of each point represents its true class. Colour indicates the true class for examples labelled by the oracle and current predicted class for those examples not yet labelled.

Figure 10: A visualisation of the active learning process running on the Reuters dataset using an uncertainty sampling selection strategy. Examples labelled by the oracle are shown enlarged. The shape of each point represents its true class. Colour indicates the true class for examples labelled by the oracle and current predicted class for those examples not yet labelled.